§ Bandits with general function approximation

Setup: For $t = 1, 2, \dots T$:

observe context $x_t \in \mathcal{X}$

take action $a_t \in A$

receive reward $r_t = f^*(x_t, a_t) + \underline{\varepsilon_t}$ 1-SG zero mean

Goal: minimize pseudo regret

$$PReg(T) = \sum_{t=1}^{T} \max_a f^*(x_t, a) - \sum_{t=1}^{T} f^*(x_t, a_t)$$

① $\mathcal{F} \subseteq (\mathcal{X} \times A \to [0,1])$

Assumption: $f^* \in \mathcal{F}$ (realizability) ② $|\mathcal{F}| < \infty$.

But functions in $\mathcal{F}$ are not nec. linear $f(x,a) \neq \langle \theta, \phi(x,a) \rangle$

Example: generalized linear bandits: linear bandits w/ an "activation fn".

$$f^* \in \mathcal{F} = \{ f_\theta(x,a) = \underbrace{\sigma}_{\text{known}}(\langle \theta, \underbrace{\phi}_{\text{known}}(x,a) \rangle) : \|\theta\|_2 \le 1 \}$$

① $$r_t \mid x_t = x, a_t = a \sim \text{Bernoulli}\left(\frac{1}{1 + e^{-\langle \theta^*, \phi(x,a) \rangle}} \right)$$

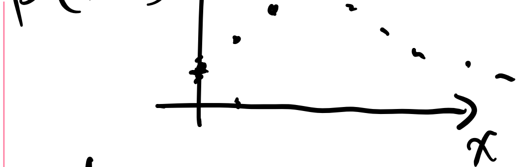
logistic bandits. (suitable for modeling binary responses, e.g. clicks) $\Rightarrow f^*(x,a) = \frac{1}{1 + e^{-\langle \theta^*, \phi(x,a) \rangle}}$

② $$r_t \mid x_t = x, a_t = a \sim \text{Poisson}\left(e^{\langle \theta^*, \phi(x,a) \rangle} \right)$$

Poisson rewards (modeling counts, e.g. number of visits) $\Rightarrow f^*(x,a) = e^{\langle \theta^*, \phi(x,a) \rangle}$

$\approx Bin(n, \frac{\lambda}{n})$ for large n.

$P(X=x)$



$\text{Poisson}(x;\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad x \in \mathbb{N}$

$\mathbb{E}[X] = \lambda.$

③ $r_t \mid x_t = x, a_t = a \sim N(\langle \theta^*, \phi(x,a) \rangle, 1) \Rightarrow f^*(x,a) = \langle \theta^*, \phi(x,a) \rangle$

④ other examples: e.g. Exponential distn,

Gamma. Pareto. weibull (key word: exponential family)

(Filippi. Cappe. Garivier. Szepesvári, 2010).

Given a nonlinear $\mathcal{F}$. how to design low-regret algorithm?

Again: optimism principle.

Algorithm (OFU: optimism in the face of uncertainty)

World model: reward model f

For $t = 1, 2, \ldots T$:

- construct confidence set $\mathcal{F}_t$ for f^*.
- observe x_t.
- Take action $a_t = \underset{a \in A}{\arg\max} \underbrace{\max_{f \in \mathcal{F}_t} f(x_t, a)}_{\text{The largest plausible reward action } a \text{ can get}}$

Again. Q1 How to construct tight $\mathcal{F}_t$. $\forall t$?

Q2 How to relate the tightness of $\mathcal{F}_t$ to alg's regret?

For Q1: can mimic linear bandits.

① define a center $\hat{f}_t$ ② let $\mathcal{F}_t$ be a ball around at $\hat{f}_t$ with certain distance metric.

① Natural idea: "best fit".

$$\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t-1} (f(x_s, a_s) - r_s)^2 =: L_{t-1}(f)$$

② How to define distance metric?

Recall: in linear bandits, cannot guarantee $\hat{f}_t$ and f^* close everywhere.

Measure closeness using "data norm".

$$S_{t-1} = \{(x_s, a_s)\}_{s=1}^{t-1}.$$

$$\mathcal{F}_t = \{ f \in \mathcal{F} : \| f - \hat{f}_t \|_{S_{t-1}}^2 \le \beta_t \}$$

Here $\|g\|_S = \sqrt{\sum_{(x,a) \in S} g(x,a)^2}$ is the "data norm" of g wrt dataset S.

So $$\mathcal{F}_t = \{ f \in \mathcal{F} : \sum_{s=1}^{t-1} (f(x_s, a_s) - \hat{f}_t(x_s, a_s))^2 \le \beta_t \}$$

shorthand: $z_t := (x_t, a_t)$, $\forall t$.

Sanity check: for linear bandits, if $f \Leftrightarrow \theta$,

$\hat{f}_t \Leftrightarrow \hat{\theta}_t$

The above $\Leftrightarrow$ $$\sum_{s=1}^{t-1} (\langle \theta - \hat{\theta}_t, \phi(x_s, a_s) \rangle)^2 \le \beta_t$$

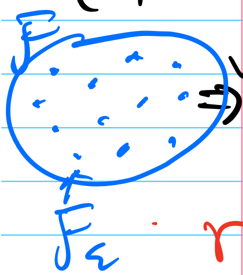
$$\Leftrightarrow \sum_{s=1}^{t-1} (\theta - \hat{\theta}_t)^T \phi_s \phi_s^T (\theta - \hat{\theta}_t) \le \beta_t$$

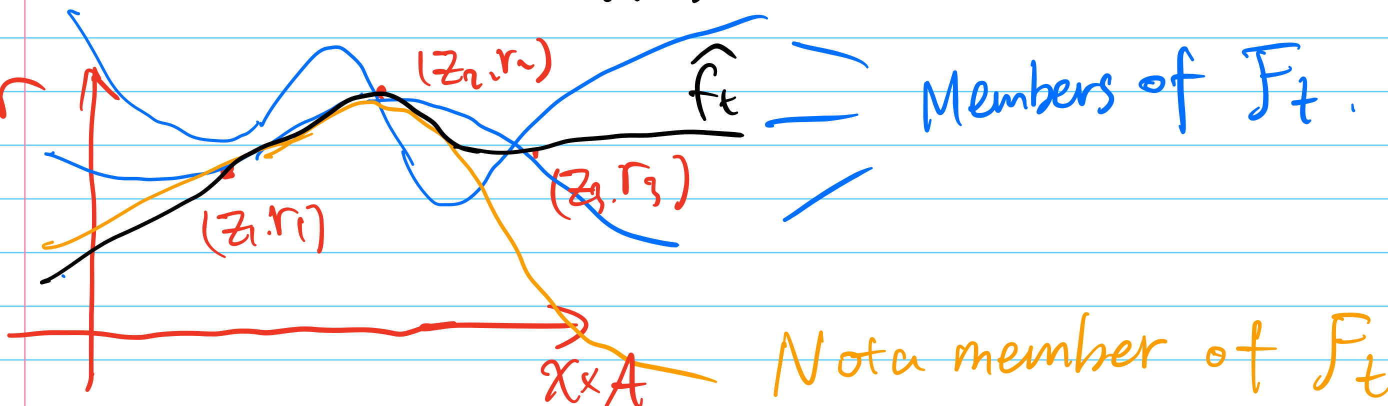
$$\Leftrightarrow \quad \|\theta - \hat{\theta}_t\|^2_{V_{t-1}(0)} \le \beta_t$$

Lemma (Validity of conf. set): let $\beta_t = 8\ln\frac{T|\mathcal{F}|}{\delta}$ $(=:\beta)$ $\forall t$, Then, w.p. $1-\delta$:

$\forall t, \quad f^* \in \mathcal{F}_t$ (validity of confidence set)

(Recall: for linear bandits, $\ln|\mathcal{F}| \approx$ # free parameters $= d$.
– For infinite $\mathcal{F}$, can consider its discretization, replace $|\mathcal{F}|$ by its covering number.
w.p. $\Rightarrow$ $\forall t.\ \|\theta - \hat{\theta}_t\|^2_{V_{t-1}(0)} \le \tilde{O}(d)$ same as before)





pf of validity lemma:

Claim: there exists event E, $P(E) \ge 1-\delta$.

on E: $\forall f \in \mathcal{F}, \ \forall t$,

$$\frac{1}{2}\|f-f^*\|^2_{S_t} - 4\ln\frac{2|\mathcal{F}|T}{\delta} \overset{①}{\le} L_t(f) - L_t(f^*) \overset{②}{\le} \frac{3}{2}\|f-f^*\|^2_{S_t} + 4\ln\frac{2|\mathcal{F}|T}{\delta}$$

Interpretation: ① If $L_t(f)$ small $\Rightarrow$ f close to f^*.

② If f is close to f^* $\Rightarrow$ $L_t(f)$ small

Claim $\Rightarrow$ lemma: on E, want to show $\|\hat{f}_t - f^*\|^2_{S_t}$

small. ($\le \beta$)

use ① or ② ? ① V

$$\Rightarrow \frac{1}{2}\|\hat{f}_t - f^*\|_{S_t}^2 - 4\ln\frac{2|F|T}{\delta} \le L_t(\hat{f}_t) - L_t(f^*) \underset{\text{optimality of } \hat{f}_t}{\le} 0$$

$$\Rightarrow \|\hat{f}_t - f^*\|_{S_t}^2 \le 8\ln\frac{2|F|T}{\delta} = \beta.$$ $\square$.

proof of the claim:

It suffices to show $\forall f, \forall t$, w.p. $1-\delta$:

$$\underbrace{\frac{1}{2}\|f-f^*\|_{S_t}^2 - 4\ln\frac{2}{\delta} \overset{ⓐ}{\le} L_t(f) - L_t(f^*) \overset{ⓑ}{\le} \frac{3}{2}\|f-f^*\|_{S_t}^2 + 4\ln\frac{2}{\delta}}_{E_{f,t}}$$

(Why? Define $E = \bigcap_{f\in F}\bigcap_{t=1}^{T} E_{f,t}$ and apply union bound (exercise))

we will prove ⓑ (ⓐ is similar)

Step 1: understand what $L_t(f) - L_t(f^*)$ concentrates to: (a r.v. should concentrate to its expectation)

$$L_t(f) - L_t(f^*) = \sum_{s=1}^{t} (f(z_t) - r_t)^2 - (f^*(z_t) - r_t)^2$$

If $z_1 \cdots z_t$ are all chosen ahead of time & deterministically. what's its expectation?

$$= \sum_{s=1}^{t} \left(f(z_s) - f^*(z_s) - \varepsilon_s\right)^2 - \varepsilon_s^2$$

the expectation here is w.r.t to the randomness of noise $\varepsilon_1 \cdots \varepsilon_t$.

$$= \underbrace{\sum_{s=1}^{t} \left(f(z_s) - f^*(z_s)\right)^2}_{\|f-f^*\|_{S_t}^2 \text{, fixed.}} - \underbrace{\sum_{s=1}^{t} 2\left(f(z_s) - f^*(z_s)\right) \varepsilon_s}_{\text{expectation zero.}}$$

Therefore, we focus on understanding the concentration of

$$L_t(f) - L_t(f^*) - \|f-f^*\|_{S_t}^2$$

$$= -\sum_{s=1}^{t} 2\left(f(z_s) - f^*(z_s)\right) \varepsilon_s =: M_t$$

$\{M_t\}$ is a martingale & $\mathbb{E}[M_t] = 0 \ \forall t$.

At this point, can use e.g. Azuma's inequality or self-normalized tail ineq. for martingales (linear bandit lecture) to bound the RHS.

However: ① this may give suboptimal bounds;

② there is a more direct way:

Step 2: utilize "subgaussian" property of M_t

Intuition: If all $z_1 \cdots z_t$ are chosen ahead,

M_t is SG with variance proxy $4\|f-f^*\|_{S_t}^2$

i.e. $\mathbb{E}\, e^{\lambda M_t - 2\lambda^2 \|f-f^*\|_{S_t}^2} \le 1 \quad \forall \lambda \quad (*)$

Fact: $(*)$ continues to be true when $z_1 \cdots z_t$ are chosen adaptively online.

(proof: use law of iterated expectations)

Fix $\lambda > 0, \delta > 0$.

Markov's ineq $\Rightarrow \lambda M_t - 2\lambda^2 \|f-f^*\|_{S_t}^2 \le \ln \frac{1}{\delta}$.

$\Rightarrow L_t(f) - L_t(f^*) \le (1+2\lambda)\|f-f^*\|_{S_t}^2 + \frac{1}{\lambda}\ln\frac{1}{\delta}$.

take $\lambda = \frac{1}{4}$: $\square$

Q2: Starting point:

Note: under E. For OFU.

$$PReg(T) = \sum_{t=1}^{T} \max_a f^*(x_t, a) - f^*(x_t, a_t)$$

same analysis as in OFUL

$$\le \sum_{t=1}^{T} \max_{f \in F_t} f(x_t, a_t) - f^*(x_t, a_t)$$

$f^* \in \mathcal{F}_t$.

$$\leq \underbrace{\sum_{t=1}^{T} \max_{f\in\mathcal{F}_t} f(x_t,a_t) - \min_{f\in\mathcal{F}_t} f(x_t,a_t)}_{\substack{\| \\ w_{\mathcal{F}_t}(z_t)}}$$

"width / uncertainty / surprise" of z_t wrt historical observations.

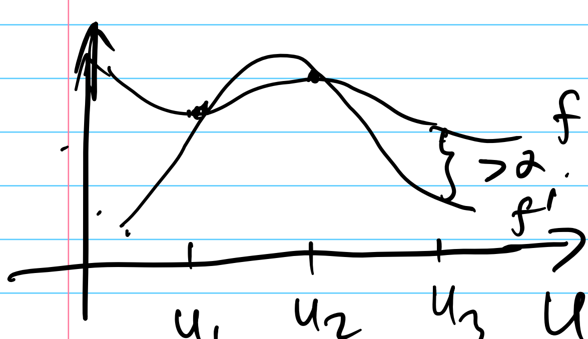
key insight: in MAB and linear bandits. $w_{\mathcal{F}_t}(z_t)$ do not stay high for a long time. due to interdependencies of the z_t's seen.

Can we characterize the interdependence of $\{z_t\}$'s for general $\mathcal{F}$?

<u>Def</u> (Eluder dimension, Russo & Van Roy, 2013).

- u_i is α-independent of $\underbrace{u_1 \cdots u_{i-1}}_{\mathcal{U}_{i-1}}$. with respect to $\mathcal{F}$. if

$\exists f, f' \in \mathcal{F}$. $\|f-f'\|^2_{\mathcal{U}_{i-1}} \leq \alpha^2$ but $|f(u_i) - f'(u_i)| > \alpha$.



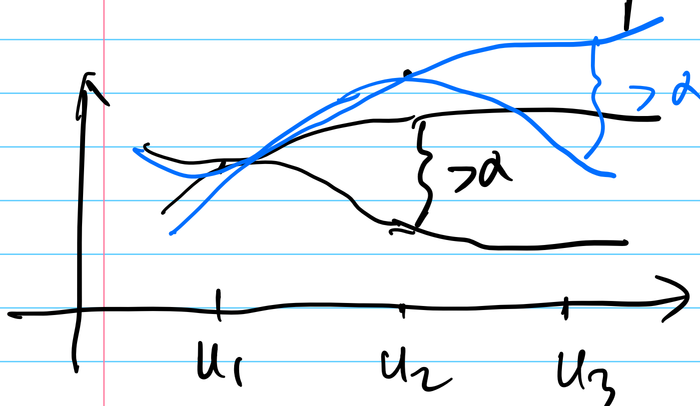
(otherwise. u_i is said to be α-dependent of $\mathcal{U}_{i-1}$

i.e. knowing f^* in U_{i-1} will imply knowledge of approximate $f^*(u_i)$.)

— The Eluder dimension of $\mathcal{F}$ at scale ε, $\mathrm{Edim}(\mathcal{F}, \varepsilon)$, is the length of the longest sequence $u_1 \cdots u_N$ $\in \mathcal{X} \times \mathcal{A}$ so that $\exists\, \alpha > \varepsilon$, and:

- u_2 is α-indept of u_1
- u_3 is α-indept of $u_1 . u_2$
- ...
- u_N is α-indept of $u_1 \cdots u_{N-1}$

call $u_1 \cdots u_N$ an α-indept sequence



... (analogy: $u_1 \cdots u_N$ is some "independent basis", similar to linear independence in linear algebra)

Politician tries to elude reporters:

— Hope to keep her true position hidden (f^*)

— Each piece of info provides needs to be new ($f^*(u_i)$) (cannot be inferred from previous ones)

— How long can she continue before her

position is determined ?

Examples of Edim:

- $\mathrm{Edim}(\mathcal{F},\varepsilon) \le |\mathcal{X} \times \mathcal{A}|$

  (an α-indept sequence cannot have repetitions)

- $\mathcal{F} = \{ \sigma(\langle \theta, \phi(x,a) \rangle) : \|\theta\|_2 \le 1, \; 0 < L_- \le \sigma'(z) \le L_+ \; \forall z \}$

  $$\mathrm{Edim}(\mathcal{F},\varepsilon) \le \tilde{O}\left(\left(\frac{L_+}{L_-}\right)^2 \cdot d \ln\frac{1}{\varepsilon} \right)$$ "well-behaved activation fn".

- (Li, Kamath, Foster, Srebro: "understanding the Eluder Dimension", 2022)

Regret analysis of OFU

Thm: w.p. $1-\delta$: $\mathrm{PReg}(T) \le O\left(\sqrt{\mathrm{Edim}(\mathcal{F},\frac{1}{T}) \cdot \ln|\mathcal{F}| T}\right)$

For generalized linear bandits:

$$\mathrm{Edim}(\mathcal{F},\tfrac{1}{T}) \le \tilde{O}\left(\left(\frac{L_+}{L_-}\right)^2 d \ln T\right) \qquad \ln|\mathcal{F}| = \tilde{O}(d)$$

$$\Rightarrow \mathrm{PReg}(T) \le \tilde{O}\left(\left(\frac{L_+}{L_-}\right)^2 d\sqrt{T}\right).$$

linear bandits: $L_+ = L_- = 1$.

proof: Claim: $\forall \alpha \geq \varepsilon$,

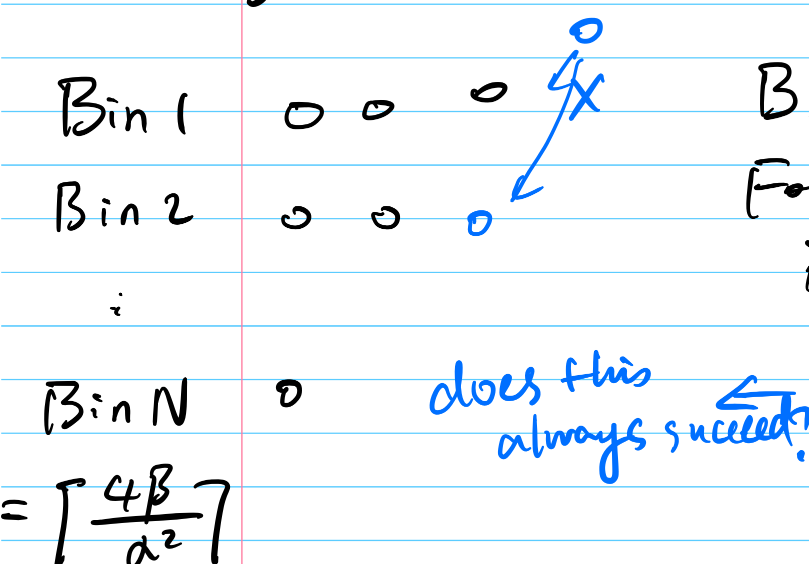
$$\sum_{t=1}^{T} \mathbb{I}\left(w_{\mathcal{F}_t}(z_t) > \alpha\right) \leq \left(\frac{4\beta}{\alpha^2} + 1\right) \mathrm{Edim}(\mathcal{F}, \varepsilon).$$

Intuition: # rounds of "surprises" should be small.

$\Downarrow$ large $w_{\mathcal{F}_t}(z_t)$

so the total surprises should also be small.

pf of Claim: we will arrange $\{z_t : w_{\mathcal{F}_t}(z_t) \geq \alpha\}_{t=1}^{T}$ in a way so that we can see its size is controlled.



$B_1 \dots B_N = [\,]$.

For $t = 1, 2, \dots T$:

  if $w_{\mathcal{F}_t}(z_t) \geq \alpha$:

    find i so z_t is α-indept of B_i

    B_i.insert(z_t)

Observation 1: the "Find" step always succeeds.

Reason: let $\bar{f}$ and $\underline{f}$ be in $\mathcal{F}_t$ s.t.

$$\bar{f}(z_t) - \underline{f}(z_t) = w_{\mathcal{F}_t}(z_t) > \alpha$$

Note: $\| \bar{f} - \hat{f}_t \|^2_{S_t} \leq \beta$

$$\| \underline{f} - \hat{f}_t \|^2_{S_{t-1}} \leq \beta$$

$$\Rightarrow \| \bar{f} - \underline{f} \|^2_{S_{t-1}} \leq 4\beta \quad \text{(exercise)}. \quad (*)$$

"Find" fails, i.e.,

If $\forall i = 1, \cdots N$, z_t is α- dept of B_i

$\Rightarrow$ for each i, $\| \bar{f} - \underline{f} \|_{B_i} > \alpha$ $\Big($ o.w. $\| \bar{f} - \underline{f} \|_{B_i} \leq \alpha$ $\underset{z_t \text{ is } \alpha\text{-dept of } B_i}{\Rightarrow}$ $\underbrace{|(\bar{f} - \underline{f})(z_t)|}_{\leq \alpha}\Big)$

$$\Rightarrow \| \bar{f} - \underline{f} \|^2_{S_{t-1}} \geq \sum_{i=1}^{N} \| \bar{f} - \bar{f} \|^2_{B_i} > \alpha^2 \cdot N \geq 4\beta.$$

contradiction w/ $(*)$ $\boxtimes$

<u>Observation 2</u>: Final $B_1 \cdots B_N$ all have size $\leq \mathrm{Edim}(\mathcal{F}, \varepsilon)$

Reason: each bin has a sequence of α-indept elts.

Therefore: $\displaystyle \sum_{t=1}^{T} \mathbb{I}(w_{\mathcal{F}_t}(z_t) \geq \alpha) = \sum_i \text{Final } B_i \text{ size}$

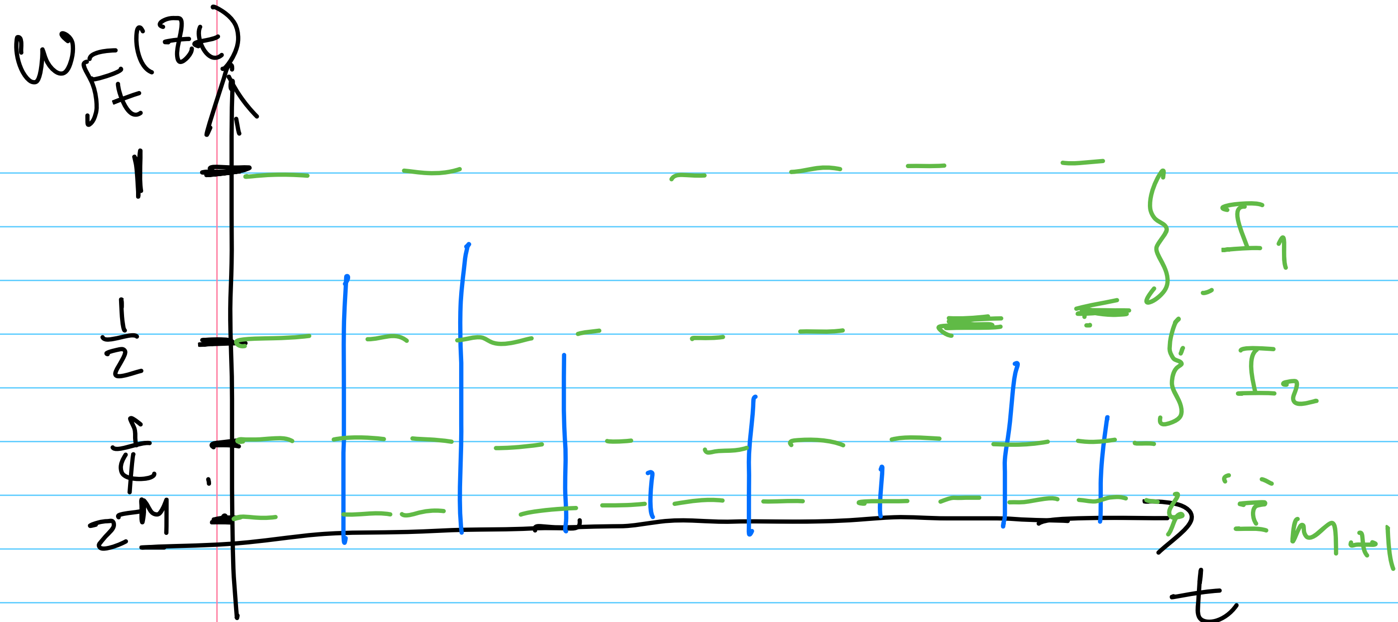
$$\leq \mathrm{Edim}(\mathcal{F}, \varepsilon) \left\lceil \frac{4\beta}{\alpha^2} \right\rceil \boxtimes$$

<u>Concluding the regret analysis</u>; partition $[T]$ by

$$I_i = \{ t : w_{\mathcal{F}_t}(z_t) \in (2^{-i}, 2^{-(i-1)}] \}.$$

$i = 1, 2, \cdots \lfloor \log \frac{1}{\alpha} \rfloor = M$ $\quad$ (let $\alpha \geq \frac{1}{T}$ TBD)

$$I_{M+1} = \{ t : w_{\mathcal{F}_t}(z_t) \leq 2^{-M} \}.$$



"peeling trick".

$$\mathrm{PReg}(T) = \underbrace{\sum_{t \in I_{M+1}} w_{\mathcal{F}_t}(z_t)}_{\le |I_{M+1}| \cdot 2^{-M} \;\le 2T\alpha.} + \underbrace{\sum_{i=1}^{M} \sum_{t \in I_i} w_{\mathcal{F}_t}(z_t)}$$

$$\le |I_i| \cdot 2^{-i+1}$$

$$\le \sum_{t=1}^{T} \mathbb{I}(w_{\mathcal{F}_t}(z_t) > 2^{-i}) \cdot 2^{-i+1}$$

$$\le (4\beta 2^{i} + 2^{-i}) \,\mathrm{Edim}(\mathcal{F}, \tfrac{1}{T})$$

$$\le 2T\alpha + \frac{8\beta\,\mathrm{Edim}(\mathcal{F}, \frac{1}{T})}{\alpha} + \mathrm{Edim}(\mathcal{F}, \tfrac{1}{T}).$$

choose $\alpha = \max\left(\frac{1}{T}, 4\sqrt{\frac{\beta\,\mathrm{Edim}(\mathcal{F}, \frac{1}{T})}{T}}\right)$

$$\leq O\Big(\sqrt{\underbrace{\beta}_{\ln|F|} E\dim(F, \tfrac{1}{T}) \cdot T} + \underbrace{E\dim(F, \tfrac{1}{T})}_{\text{lower order}}\Big).$$

$\cancel{D}$